

## “VIRUS HUNTING” USING RADIAL DISTANCE WEIGHTED DISCRIMINATION<sup>1</sup>

BY JIE XIONG, D. P. DITTMER AND J. S. MARRON

*University of North Carolina at Chapel Hill*

Motivated by the challenge of using DNA-seq data to identify viruses in human blood samples, we propose a novel classification algorithm called “Radial Distance Weighted Discrimination” (or Radial DWD). This classifier is designed for binary classification, assuming one class is surrounded by the other class in very diverse radial directions, which is seen to be typical for our virus detection data. This separation of the 2 classes in multiple radial directions naturally motivates the development of Radial DWD. While classical machine learning methods such as the Support Vector Machine and linear Distance Weighted Discrimination can sometimes give reasonable answers for a given data set, their generalizability is severely compromised because of the linear separating boundary. Radial DWD addresses this challenge by using a more appropriate (in this particular case) spherical separating boundary. Simulations show that for appropriate radial contexts, this gives much better generalizability than linear methods, and also much better than conventional kernel based (nonlinear) Support Vector Machines, because the latter methods essentially use much of the information in the data for determining the shape of the separating boundary. The effectiveness of Radial DWD is demonstrated for real virus detection.

**1. Introduction.** A current major scientific challenge is the detection of viruses in human blood samples. Cogent examples include HIV, the cause of AIDS; poliovirus, which was considered eradicated, but has now emerged in Syria and the Middle East; or middle east respiratory syndrome (MERS), which entered the United States in May 2014 via a 44-year-old male who traveled from Jeddah, South Africa, to Orlando, Florida, via London. At home he developed fever, chills and a slight cough. He was admitted to the hospital and later diagnosed with the MERS coronavirus. Since May 9, 2014,

---

Received May 2014; revised August 2015.

<sup>1</sup>Supported by public health service Grants CA019014, and AI107810 to DPD.

*Key words and phrases.* Virus hunting, nonlinear classification, high-dimension low-sample size data analysis, DNA sequencing.

This is an electronic reprint of the original article published by the  
 Institute of Mathematical Statistics in *The Annals of Applied Statistics*,  
 2015, Vol. 9, No. 4, 2090–2109. This reprint differs from the original in pagination  
 and typographic detail.

the World Health Organization (WHO) reported 536 laboratory-confirmed cases of MERS, including 145 deaths [WHO (2014)].

For an effective treatment, a rapid and accurate detection of the source of viral infection is crucial. The recent advent of deep DNA sequencing techniques has led to a potentially powerful approach and it gives rise to a new type of classification (discrimination) challenge.

A useful data space for virus detection comes from the DNA sequence and alignment process, where virus-positive (the +1 class) and virus-negative (the -1 class) samples are sequenced and the sequenced DNA reads from each sample are aligned to a target virus. Reviews of the DNA-sequence techniques can be found in Goldstein et al. (2013), Mwenifumbo and Marra (2013), Rehm (2013) and Grada and Weinbrecht (2013). A data vector counting the number of reads aligned to each nucleotide position on the target virus is obtained for each sample. Thereafter, data vectors from the 2 classes form the training set.

Because DNA sequencing is done at the base pair level of resolution, the read depth vectors are quite long. As the number of samples is relatively smaller, this analysis lies in the domain of high-dimension low-sample size (HDLSS) data, which is an active research area where the dimension  $d$  of the data vectors is larger than the sample size  $n$ ; see, for example, Hall, Marron and Neeman (2005), Liu et al. (2008), Jung and Marron (2009), Fan and Lv (2010), Shen, Shen and Marron (2013) and Yata and Aoshima (2013). In this paper we focus on HDLSS binary classification problems; see Marron, Todd and Ahn (2007) and Jiang, Marron and Jiang (2009) for some examples.

An important aspect of DNA sequencing is that increasing the total number of reads generated from a sample inflates entries of a data vector. In many cases, the number of reads generated from samples differs either due to the sequencing platform or to experimental settings; see, for example, in Metzker (2010). The fact that we usually collect different numbers of reads for the samples is regarded as a *bench effect* here. Bench effects may negatively impact the classification and need to be handled properly. This is done here by normalizing each data vector by dividing the entries by the  $L1$  norm, which is simply the summation of the entries in that vector, since they are nonnegative. Therefore, normalized vectors all have unit  $L1$  norms to control for the bench effect.

Instead of normalizing by the  $L1$  norm, other methods could be used to adjust for the bench effect, for example, normalizations based on the companion human genome. Such methods would be attractive in situations where the goal was to determine the amount of virus present. However, here our goal is to determine the presence or absence of virus (and, in particular, we are focusing on trying to find rather small amounts). In this context, our normalization seems the most powerful.

A consequence of this L1 normalization is that the normalized data vectors can be geometrically represented as points on the standard unit simplex. Data points with more nonzero entries lie more toward the interior of the unit simplex. When all entries are approximately the same, the data point is near the center. On the contrary, the more zeros in a vector, the closer this data point is to one of the vertices of the unit simplex. In the extreme case with only one nonzero entry “1” in the vector, the data point is at a vertex.

Figure 1 shows how different the virus positive and virus negative samples are, and motivates exploiting simplex geometry, by showing an overlaid plot of normalized data vectors from an HSV-1 (a human herpesvirus) detection problem. HSV-1, or human herpesvirus-1, is the leading cause of nontraumatic blindness and can cause fatal encephalitic disease in children. The virus can be treated with acyclovir, if and only if diagnosed rapidly and accurately. Both serum and cerebral spinal fluid are used for diagnosis and can be readily obtained for sequencing. In Figure 1, we overlaid 2 (out of 8) data vectors for the HSV-1 positive (the +1) class (top panel) and

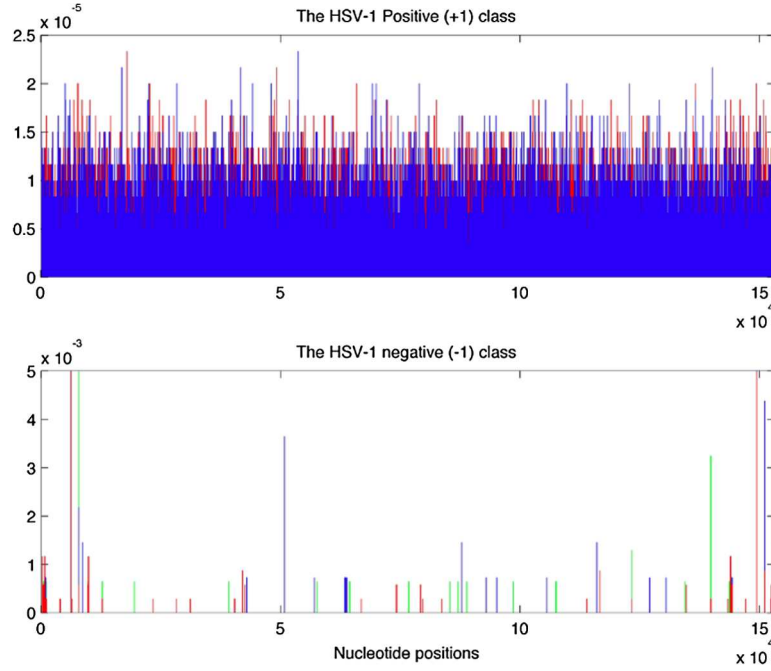


FIG. 1. Overlaid plot of 2 normalized data vectors from the HSV-1 positive (the +1) class in the top panel and 3 data vectors from the HSV-1 negative (the -1) class in the lower panel, all with different colors. The overall entries of the +1 data vectors are relatively small and have quite comparable amplitudes, while the entries of the -1 data vectors have “spikes” (which are located at quite divergent positions).

3 (out of 24) data vectors from the HSV-1 negative (the  $-1$ ) class (lower panel). The overall entries of the positive data vectors are relatively small and have relatively comparable amplitudes (top panel). The nonzero entries of the negative data vectors are very sparse and have much larger amplitudes (about 200 times larger than that of the positive samples, lower panel of Figure 1). This is a property of all virus detection problems, since the negative sequences are chosen to be genetically very different from the virus. The aligned reads (from the negatives to the virus sequence), on the other hand, are often short stretches of sequence which are of reduced complexity, that is, repeats or single nucleotide (either A, C, T or G) runs.

If we keep the same range of  $y$ -axis in plotting the positive data vectors as in plotting the negative data vectors, one can see almost nothing since the amplitudes of the former ones are much smaller than the latter ones. Equivalently speaking, the positives are close to the center of the simplex, while the negatives lie near to a diverse set of vertices of the unit simplex, because the differently colored spikes of the negative data vectors are located at quite divergent positions. A simple model for data on the unit simplex is given in Figure 2, where the 3- $d$  unit simplex is shown as a gray triangle while some  $+1$  ( $-1$ ) class data are shown as red plus signs (or blue circles, resp.). It is not hard to see that linear methods will struggle to capture the differences between classes in this case.

In our work, the DNA alignment data vectors are often of dimension 100,000 to 200,000, while the sample size is usually much smaller. Equivalently, data can be seen as points on the high-dimensional unit simplex.

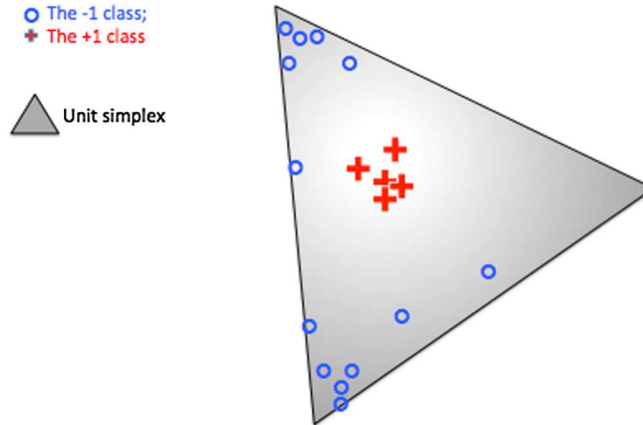


FIG. 2. A simplified example of normalized data vectors of the positive ( $+1$  class) data (red plus signs) and the negative ( $-1$  class) data points (blue circles). The unit simplex is shown as the gray triangle. Because there are many zeros in the  $-1$  data vectors, they typically locate at the vertices of the unit simplex while the positives are closer to the center.

Figures 1 and 2 suggest that the  $-1$  class departs from the center of the simplex (where the  $+1$  class lies) in many diverse directions so that the theoretical Bayes classification boundary (assuming a probability distribution for each class) is highly nonlinear. Note that the discrimination in radial directions appears to be attractive. This motivates the development of Radial DWD in order to incorporate such a nonlinear pattern. As detailed in Section 4, by optimizing a hypersphere over its center and radius, Radial DWD separates the 2 classes, favoring putting the  $+1$  ( $-1$ ) class inside (outside) the hypersphere. The computation of Radial DWD through solving a sequence of Second Order Cone Programs Alizadeh and Goldfarb (2003) is carried out by an interior point optimization package called SDPT3, developed by Tutuncu, Toh and Todd (2001). A future sample will be classified as  $+1$  ( $-1$ ) when it is located inside (outside) the hypersphere.

A standard approach to HDLSS classification problems is linear methods, such as Mean Difference [MD, Schölkopf and Smola (2002)], penalized logistic regression with LASSO penalty [LASSO, Tibshirani (1996)], Support Vector Machine [SVM, Vapnik (1995), Shawe-Taylor and Cristianini (2004)] and Distance Weighted Discrimination [DWD, Marron, Todd and Ahn (2007)]. Figure 2 suggests that, as the dimension and diversity of the  $-1$  class grow, such methods will be severely inefficient. This issue is carefully studied for actual “virus hunting” in Section 2 and by simulation in Section 3. It is natural to wonder if a more serious competitor to Radial DWD is a nonlinear kernel Support Vector Machine classification [Burges (1998), Hastie, Tibshirani and Friedman (2009)]. The most popular of these is the Radial Basis Function (RBF) kernel. The virus detection capability of these methods are compared in Figure 3, where RBF kernel SVM and Radial DWD classification are illustrated. Note that in the machine learning literature, RBF is a synonym for “Gaussian kernel.”

In Figure 3, RBF kernel SVM and Radial DWD are trained using 8 HSV-1 positive (red plus signs) and 24 HSV-1 negative (blue circles) data vectors, which are partially shown as an overlaid plot in Figure 1. Signed distances to the corresponding separating boundary (the black vertical dashed line) are depicted along the  $x$ -axis as a jitter plot. Random heights are assigned in order to visually separate the points. Additionally, 127 new samples are used as a test set and kernel density estimates are given for each group. While the majority of test samples are shown as gray  $x$ -symbols, 4 are highlighted in magenta since they are HSV-1 positive human samples; 14 are highlighted in green since they are highly related herpesviruses (with nonhuman hosts). The related viruses share significant sequence identity (traditionally larger than 35%) with the reference virus, but may infect animals rather than humans. Domestic cats and cattle, for instance, can be infected with a herpesvirus homologous to HSV-1.

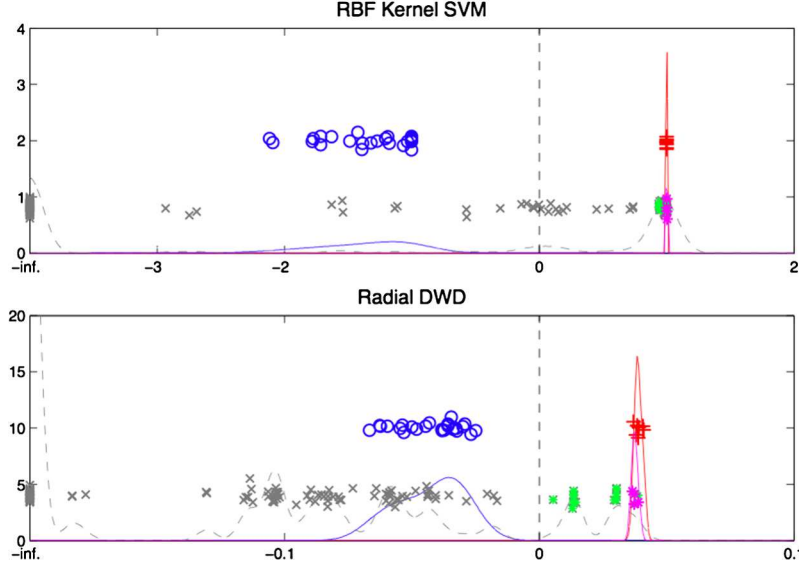


FIG. 3. An HSV-1 classification example to compare the performance of RBF kernel SVM and Radial DWD. Trained on red plus signs versus blue circles, the former method endures a high false positive error since many negative test samples (gray x-symbols) are on the same side of the separating boundary as the positive class, while Radial DWD successfully classified all positive HSV-1 samples (magenta asterisks) and related viruses (green asterisks) with no false positive.

The performance of RBF kernel SVM is far from satisfactory: although positive samples (magenta and green asterisks) are very close to the true positives, many (69) grays (unrelated samples) are also classified as HSV-1 positive. This is expected since kernel methods require a type of “data richness,” that is not present in the virus hunting problem. In particular, they work well in situations where training data can be found in all of the various regions where the test data will appear. But in virus hunting data analysis, that completely breaks down.

Radial DWD shows a superior classification result not only because it correctly classified all HSV-1 positive samples but also because the positive samples are grouped reasonably well: HSV-1 positive human samples (magenta asterisks) are tightly clustered with the positive training data (red plus signs); related herpesviruses (green asterisks) are clustered according to the host species that they infect—from the right to the left—monkey, pig and cattle. The grouping property of Radial DWD can be exploited to classify new viruses, for example, in different animal hosts, as they would be related, but not identical to the known ones.

For some data sets, it will be sensible to use a given point, for example, the center of the simplex or the sample mean, as the centerpoint of the sep-

arating sphere. Therefore, solving the associated optimization problem will be generally easier. However, the center of the simplex seems inappropriate for virus hunting, as due to various biological effects, even in the limit as the number of reads goes to infinity, the read depth vector is not flat. The sample mean can be appropriate in many situations, but as the centroid classifier is often a lot less efficient in many high-dimensional biological settings, we expect Radial DWD to often be worth the overhead of the more complex optimization. Furthermore, we also have our eye on generalizing to other data types, where we believe the property of Radial DWD having conventional DWD as a limit (as the center goes to infinity in a particular direction, with the radius also growing) will become very important.

A full description of this HSV-1 classification is given in Section 2, where we carefully compare Radial DWD with some linear and nonlinear competitors and the superiority of Radial DWD under this radial context is discussed in detail. A similar conclusion can be drawn from the simulation study in Section 3. Radial DWD optimization and an iterative algorithm to solve it can be found in Section 4. An introduction to virus detection, insights about the Dirichlet distribution and the high-dimensional unit simplex, along with more details of our data sets and some proofs, can be found in the supplementary materials in Xiong, Dittmer and Marron (2015).

**2. Virus detection data analysis.** As briefly described in Section 1, Radial DWD presents an appealing virus detection capability. A broader comparison between Radial DWD and its linear and nonlinear competitors is given in this section through analyzing a real data example of detecting the  $\alpha$ -Human Simplexvirus 1 ( $\alpha$ -HSV-1 or HSV-1). This virus is a subfamily of Human Herpesvirus (HHV). The data set consists of the following 2 subsets:

- The training data are  $n_+ = 8, n_- = 24$  vectors of dimension 152,261, which is the DNA length of HSV-1. Entries of each data vector correspond to the nucleotide positions in the virus DNA sequence. The training data of the +1 (HSV-1 positive) and -1 (HSV-1 negative) classes are normalized to the unit simplex (of dimension 152,261). The +1 class tends to locate near the center while the -1 class tends to locate near a diverse set of vertices of the simplex. Classifiers are trained using the +1 versus the -1 classes.
- The test set consists of the DNA alignment vectors from the following samples: 4 HSV-1 positive human samples (not appearing in the training), 14 nonhuman  $\alpha$  Simplexvirus-1 (including 5 monkey Simplexvirus-1, 8 pig Simplexvirus-1 and 1 cow Simplexvirus-1) and 109 much more distantly related viruses.

Nonzero data vectors are normalized to the unit simplex and can be viewed as points on it. Samples whose DNA alignment vectors are zero



vectors are put at  $-\infty$ . This is reasonable since zero vectors only exist in the  $-1$  class training set or the test set: (a) if the sample comes from the  $-1$  class training set, it has no effect on the calculation of the separating sphere (interpreting the reciprocal of  $-\infty$  to be zero); (b) if the sample comes from the test set, it should surely be classified as  $-1$  and  $-\infty$  is viewed as outside the separating hypersphere.

The classification performance of Radial DWD is compared with a number of popular classification methods including MD, LASSO, linear DWD, SVM and RBF Kernel SVM in Figure 4. Quadratic SVM gave results that were quite similar to RBF SVM, so it is not shown here. For methods including a separating plane, relative performance comes from the projection onto the normal vector, shown as the horizontal axes in Figure 4. Radial DWD is similarly interpreted as the signed distance to the separating sphere. The  $+1$  training data are shown as red plus signs,  $-1$  training data as blue circles, HSV-1 positive human samples (real human DNA samples that are infected by HSV-1) as magenta asterisks, related  $\alpha$  simplex herpesviruses as green asterisks and other samples as gray  $x$ -symbols (known to be HSV-1 negative). The position of the separation boundary is shown by the black vertical dashed line. Signed distances to the separating boundaries are de-

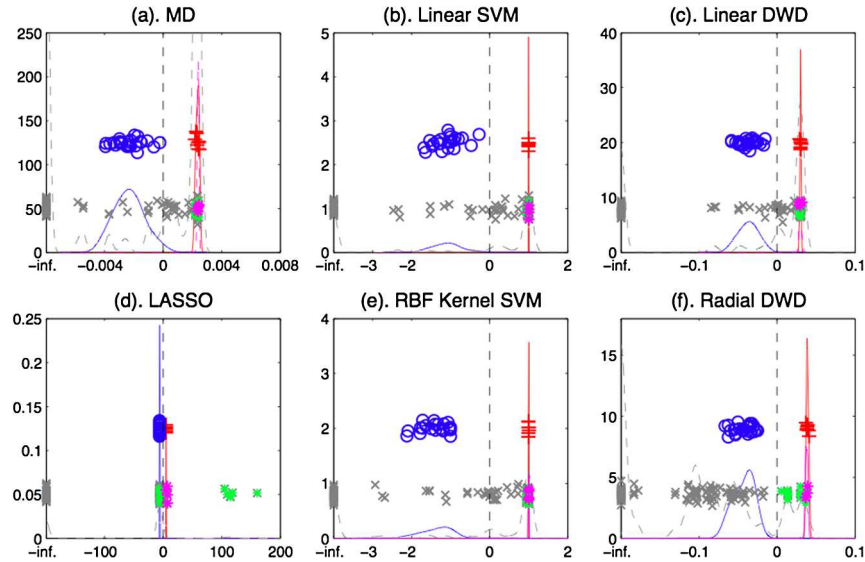


FIG. 4. Real data example of an HSV-1 classification problem. We show 6 panels of 1-dimensional “signed distance to separating boundary” plots to compare Radial DWD [panel (f)] with MD, LASSO, linear DWD, SVM and RBF SVM. Red plus signs are  $+1$ , blue circles are  $-1$ , magenta asterisks are HSV-1 positive humans, green asterisks are related nonhuman herpesviruses, gray  $x$ -symbols are nonpositive samples. Figure 4 shows the superior performance of Radial DWD.



picted along the horizontal axis, while the vertical perturbation is used for visual separation of the points. Kernel density plots are provided as well.

Panels (e) and (f) are the same as in Figure 3 (except that the length-width ratio of the figures are different) and the superior performance of Radial DWD is explained there. While the training data is well separated in all cases in Figure 4, the good classification property may not carry over to the test samples. The performance of MD, SVM, DWD and RBF SVM tend to be similar in this example where the false positive rates are very high (larger than 50%), that is, most of the negative gray  $x$ -symbols are to the left of the dashed line. This performance contrasts sharply with panel (f) where all gray  $x$ -symbols are to the right. Meanwhile, LASSO presents a unique behavior with zero false positive. However, it fails to correctly classify 8 (out of 14) HSV-1 related viruses (green asterisks) since they fall on the left-hand side of the LASSO-separating hyperplane. The other 6 HSV-1 related viruses are much further from the positive training data (red plus signs) to the right. Our simulations show that LASSO tends to pick out a small subset of nucleotide positions and classify data merely based on very limited information gained on those positions, which results in poor classification.

An additional insightful comparison of methods using simulated data sets appears in Section 3 and Radial DWD will be shown to have a much better classification accuracy in terms of both lower false positive and lower false negative error rates, while all the other competitors considered here perform poorly. Note that real data examples of  $\beta$ -HHVs and  $\gamma$ -HHVs (the other 2 subfamilies of HHV) classification were also analyzed and examples can be found in the supplementary materials in Xiong, Dittmer and Marron (2015).

In addition to giving outstanding classification results when one class is widely distributed around the other, the computation of Radial DWD is fast enough to be useful for modern scale bioinformatics data sets. The computational speed is nearly independent of the dimension of the data vectors because the method is based on a QR decomposition (see Section 4 for detail). In particular, the full set of simulations shown in Section 3, involving many replications, was done in a few hours.

**3. Simulation study.** Section 2 showed that Radial DWD outperforms its linear and nonlinear competing classifiers for real virus detection data, and this idea is further emphasized in this section by a simulation study. Our simulations are based on Dirichlet distributions which are a popular and broad family of distributions on the unit simplex. Figure 5 shows the classification results, aimed at modeling the behavior observed in real data in Figure 4, detailed in Section 3.1. Broader simulation results are discussed in Section 3.2.

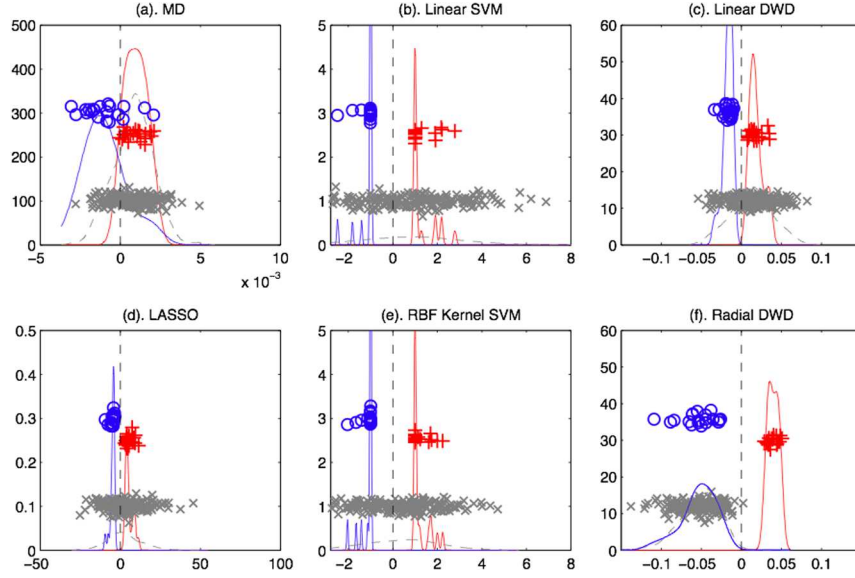


FIG. 5. A simulated example illustrating the potential for improved performance of Radial DWD, in the spirit of Figure 4. Class +1 is shown as red pluses, -1 as blue circles, test samples as gray  $x$  symbols. The vertical axis shows random heights to visually separate the points, along with kernel density estimates (i.e., smooth histograms). The separating boundaries are calculated using the following: (a) MD, (b) Linear SVM, (c) Linear DWD, (d) LASSO, (e) RBF SVM and (f) Radial DWD. Except Radial DWD, all the other methods have poor classification performance for the test samples, which should be mostly to the left of the dashed line in each case.

3.1. *Simulation 1.* The simulated data in Figure 5 have dimension  $d = 50$ , with  $n_+ = 20$  class +1 samples represented as red plus signs and  $n_- = 20$  class -1 samples represented as blue circles. Data are simulated using the Dirichlet distribution  $\text{Dirichlet}(\alpha)$ , supported on the unit simplex.

The parameter  $\alpha \in R^{d+}$  determines the mode and dispersion of the Dirichlet distribution. If all the entries in  $\alpha$  are the same, the distribution is centered on the unit simplex. Suppose the common entries are larger (less) than 1, increasing (decreasing) the entries makes the distribution more concentrated to the center (vertices) of the unit simplex; suppose the common entries are exactly 1s, the corresponding distribution is the uniform on the simplex. Examples in 3 dimensions can be found in the supplementary materials in Xiong, Dittmer and Marron (2015).

The +1 class data in Figure 5 are drawn from  $\text{Dirichlet}(\alpha_+)$  with  $\alpha_+ = (5, \dots, 5)$  and the -1 class data are generated from  $\text{Dirichlet}(\alpha_-)$  with  $\alpha_- = (0.5, \dots, 0.5)$ .

Classifiers, including MD, LASSO, linear SVM, RBF Kernel SVM, DWD and Radial DWD, are trained on the red pluses and blue circles. We assess

the performance by classifying 200 new test samples drawn from the  $-1$  class population. The test samples are shown in Figure 5 as gray  $x$  symbols. Note that the Quadratic Kernel SVM (QSVM) was also considered. It performed very similarly with RBF SVM in this particular example, and hence is not shown here.

In each panel of Figure 5, the signed distances of the data points to the optimal separating hyperplane are shown on the horizontal axes. The position of each separating hyperplane is shown as a dashed line. Data points that fall on the same side of the hyperplane as the  $+1$  ( $-1$ ) class will have positive (negative) distances. Kernel density plots (e.g., smooth histograms) are provided as another way of viewing the population of each class. As shown in Figure 5(a), MD performs poorly (with many gray test points to the right of the boundary) since the separation of classes in this example is not a shift of means. In particular, 152 out of 200 samples are misclassified as  $+1$ . Figure 5(b) shows that the 2 training classes are linearly separable by using SVM, but the training data from both classes pile up at the margin. Moreover, 129 out of 200 test samples are misclassified as  $+1$ .

Data piling is a sign of overfitting and is very undesirable since the corresponding separating hyperplane is driven heavily by the particular realization of the data at hand [see Marron, Todd and Ahn (2007)]. DWD was developed to address this ubiquitous problem with SVM, yet Figure 5(c) is similar to (b). The phenomenon of data piling is diminished as expected from the ideas of Marron, Todd and Ahn (2007). However, the performance of DWD for this test set is far from satisfactory because radial separation is the key: again, many (142 out of 200) test samples are misclassified as  $+1$ .

Classification using LASSO is illustrated in Figure 5(d). The training data are well separated, but 117 of the 200 test samples are misclassified. Figure 5(e) shows the classification using the RBF (nonlinear) Kernel SVM. When the training set is linearly separable, kernel SVM behaves like the linear counterpart but may overfit the training data more severely under HDLSS assumptions. Although the dimension is fairly moderate, data piling still exists. The expected improvement over the linear counterpart is present in the sense that only 120 out of 200 test samples are misclassified, although this is still unacceptably poor.

A much improved performance and classification accuracy can be observed in Figure 5(f) where Radial DWD is applied. Training data are well separated with no signs of data piling and, except for one test data point, all the other test samples are correctly classified, showing that Radial DWD solves the overfitting problem one may intuitively expect from the RBF kernel SVM in HDLSS radial contexts.

It is not surprising that, despite the underlying nonlinear pattern, SVM and DWD successfully separate the 2 training classes due to the large size of the data space. However, the good classification performance does not carry

over to the test samples, which may differ from the +1 class in directions that do not appear in the  $-1$  class training data. This highlights the limitation of linear methods in this type of context. Figures 4 and 5 together make it clear that the intuitive ideas in Section 1 are indeed the drivers of the observed superior performance of Radial DWD. Thus, simulating data from the Dirichlet distribution is useful and insightful to understand the data structure of the virus discovery.

*3.2. Simulation 2.* Next a broader simulation study is conducted. The training data and the test data are simulated on the unit simplex using  $\text{Dirichlet}(\alpha)$  with  $\alpha$  summarized in Table 1. In each example,  $n_+ = 20$  +1 class and  $n_- = 50$   $-1$  class data of dimension  $d = 10, 50, 100, 500, 1000, 5000, 10,000, 50,000, 100,000$  are generated in order to cover a range from non-HDLSS to extreme HDLSS cases. Additionally, in panels (a1) and (a2), 5000 test samples are drawn from the  $-1$  class in order to assess the false positive rate; in panels (b1) and (b2), 5000 test samples are drawn from the +1 class in order to assess the false negative rate. Thirty repetitions are done for each case and each dimension.

The tuning parameters in LASSO, (linear/Quadratic/RBF) SVM, DWD are determined by 5-fold cross-validation. Classifiers are trained using the +1 versus the  $-1$  class. Classification error (false positive and false negative) is calculated for classifying the 5000 test samples and is illustrated in Figure 6.

In the first simulation in panels (a1), (b1) and (c1), the +1 class is simulated uniformly on the simplex using  $\text{Dirichlet}(1 \cdots 1)$ , while the  $-1$  class is simulated near the vertices of the simplex, as given in Table 1. The class separation is hard in low dimensions, but, as dimension grows, the relatively low sample size of the training data makes the separation easier. It can be seen in panel (a1) that when dimension is low (around 10), RBF kernel SVMs and Radial DWD perform similarly well with false positive error rates below 10%, LASSO and Quadratic kernel SVM follows and all the other linear methods perform poorly. As dimension goes to  $\infty$ , the false positive error of Radial DWD shrinks to zero quickly, while that of the MD/SVM/DWD/RBF kernel SVM/Quadratic kernel SVM goes to 1; that of LASSO converges to around 50%. A quite different tendency can be observed in panel (b1) when the false negative rate is being examined. When dimension is low, LASSO tends to have a very large false negative error, but the error shrinks to zero quickly as dimension grows, as do the false negative error rates for the other methods. The average of the 2 types of errors is summarized in (c1). It is not hard to see that the kernel SVMs and Radial DWD are comparably good in low dimensions; the error rate of the former one converges to around 50%, while that of the latter one converges to zero quickly as dimension grows. Additionally, the average error rate of

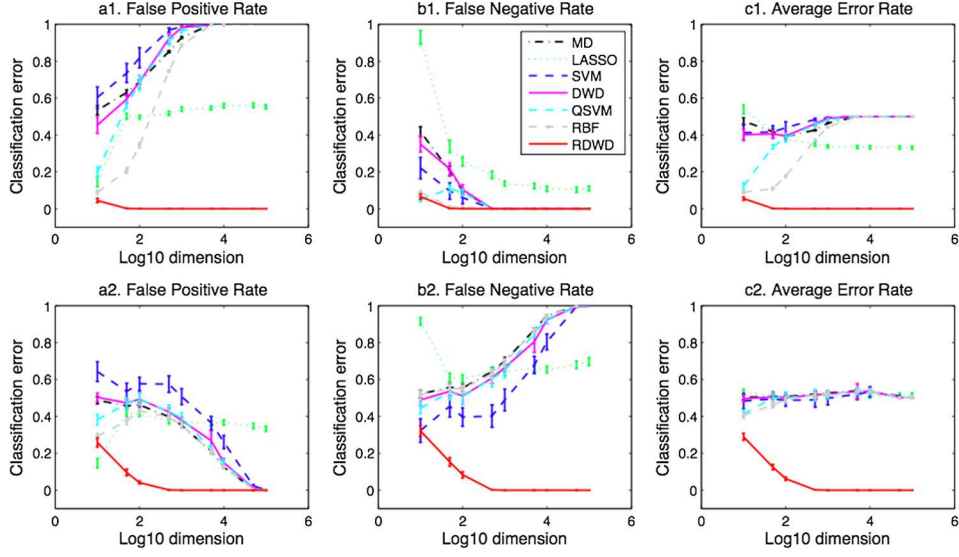


FIG. 6. A simulation study illustrating the potential for improved performance of Radial DWD. The false positive rate is depicted in panels (a1) and (a2) under each parameter setting, with the corresponding false negative rate (under the same training setup) in (b1) and (b2). The average of the false positive and false negative rate is shown in panels (c1) and (c2), respectively. Classification error is calculated for the following: — MD — LASSO, — Linear/Quadratic/RBF SVM, — Linear DWD and — Radial DWD (RDWD). A color key is also given. Error bars are obtained by repeating the simulation 30 times for each dimension  $d$ . Figure 6 shows the outstanding performance of Radial DWD relative to typical methods in these radial settings.

MD/SVM/DWD is relatively stable (around 50%); the average error rate of LASSO is around 35% for large dimensions.

The second simulation in panels (a2), (b2) and (c2) is similar to the first except that the  $-1$  class is closer to the center. This is even a harder classification problem when dimension is low. An almost opposite tendency could be observed in (a2) and (b2), compared to (a1) and (b1). Except LASSO, the false positive rate [in (a2)] of all methods shrinks to zero, while that of Radial DWD decreases much faster; the false positive error rate of LASSO is around 35% for large dimensions. Shown in panel (b2), the false negative

TABLE 1  
Parameter  $\alpha$  used in simulation

Case #	+1 class	−1 class	Corresponding panels in Figure 6
1	(1...1)	(0.1...0.1)	(a.1) (b.1) and (c.1)
2	(1...1)	(0.5...0.5)	(a.2) (b.2) and (c.2)

error rate of Radial DWD still decreases to zero as dimension grows, however, the error rates of its competitors goes to 1 (or above 60% for LASSO). The average of the false positive and the false negative rate is illustrated in panel (c2) where a similar pattern as (c1) can be observed, except that even in low dimensions, kernel SVMs did not work as well as Radial DWD. When dimension is high, all Radial DWD’s competing classifiers have error rates around about 50% (i.e., essentially random choice).

We also studied several other examples [see Supplement in Xiong, Dittmer and Marron (2015)]. They show fairly similar results. As suggested by our current simulations, Radial DWD outperforms MD, LASSO (linear, Quadratic, RBF) SVM and linear DWD when the radial separation is the key player to discriminate classes. As noted before, the full set of simulations shown in Section 3, involving many replications, was done in a few hours.

Radial DWD performs well with this type of data because of the particular geometry. In Figure 1, we show that there are scaling issues with these coverage vectors as data objects, which are handled by dividing by the sum of the entries. This transformation means the data live on the unit simplex, hence, we study its geometry. Furthermore, because the dominant spikes in Figure 1 are in different locations, the data negative samples are widely distributed around the simplex, in many different directions. We tried to illustrate this phenomenon with a grossly simplified (because human perception tends to fail beyond 3 dimensions) toy example in Figure 2. But it is the major exaggeration of this effect, that naturally occurs in this HDLSS context, that drives the major breakthrough of Radial DWD relative to the existing competitors (which were not designed for this setting).

#### 4. Radial DWD optimization.

4.1. *Formulate the optimization problem.* To set notation, let  $n$  denote the number of training  $d$ -vectors  $x_i$  with corresponding class labels  $y_i \in \{-1, +1\}$ ,  $i = 1 \cdots n$ . We let  $X$  denote the  $d \times n$  matrix whose columns are  $x_i$ . Let  $e$  denote an  $n$ -vector of 1s. Let  $n_+ = \sum_{i=1}^n I_{\{y_i=+1\}}$  and  $n_- = \sum_{i=1}^n I_{\{y_i=-1\}} = n - n_+$  be the sample size of the  $+1$  class and the  $-1$  class, respectively. Denote  $O \in R^d$  as the center of a candidate separating sphere, and let  $R \in R^+$  be the radius, and define the signed residual of the  $i$ th data point as  $\bar{r}_i = y_i(R - \|x_i - O\|_2)$ , where  $\|\cdot\|_2$  represents the Euclidean norm. We would like to search for  $O$  and  $R$  such that  $\bar{r}_i$  are positive and large, which requires the  $+1$  class to lie inside and the  $-1$  class to lie outside the hypersphere. However, in order to incorporate the case when the 2 classes are not separable by a hypersphere, we allow classification error by adding nonnegative “slack variable”  $\varepsilon_i$ , as in Burges (1998) and Marron, Todd and

Ahn (2007), and define perturbed residuals as  $r_i = y_i(R - \|x_i - O\|_2) + \varepsilon_i$ . We now define the optimization problem for Radial DWD as follows:

$$\begin{aligned}
 (4.1) \quad & \text{Min}_{r,O,R,\varepsilon} \sum_i \frac{1}{r_i} + Ce^T \varepsilon \\
 \text{s.t.} \quad & r_i = y_i(R - \|x_i - O\|_2) + \varepsilon_i, \quad i = 1, \dots, n, \\
 & R \geq 0, \quad r \geq 0, \quad \varepsilon \geq 0,
 \end{aligned}$$

where  $r$  is a vector of  $r_i$ ,  $\varepsilon$  is a vector of  $\varepsilon_i$  and  $r \geq 0$  and  $\varepsilon \geq 0$  are in the component-wise sense, and  $C$  is the penalty parameter of misclassification, as appears in SVM and DWD. It can be seen that the influence of the  $-1$  class data decreases as they get further away from the separating hypersphere. The influence shrinks to zero for the  $-1$  class data located at infinity. However, this is not true for the  $+1$  class (because of the penalty term). Following Marron, Todd and Ahn (2007), we linearize the objective function by defining  $\rho_i = (r_i + \frac{1}{r_i})/2$  and  $\sigma_i = (\frac{1}{r_i} - r_i)/2$ , so that  $\frac{1}{r_i} = \rho_i + \sigma_i$ ,  $r_i = \rho_i - \sigma_i$ . Additionally, we relax the constraints

$$\{\rho_i^2 - \sigma_i^2 = 1, \rho_i - \sigma_i \geq 0, i = 1, \dots, n\}$$

to the second order cone constraint

$$\{(\rho_i, \sigma_i, 1) \in S_3, i = 1, \dots, n\},$$

where the Second Order Cone of dimension  $k$  is defined as

$$S_k = \{(\varsigma; \mu) \in R^k : \varsigma \geq \|\mu\|_2\}.$$

One can show that when the 2 classes are separable by using a hypersphere, this relaxation will not change the optimal solution. By the transformation of  $\frac{1}{r_i}$  and the substitution with Second Order Cone constraints, the optimization problem becomes

$$\begin{aligned}
 (4.2) \quad & \text{Min}_{\rho,\sigma,O,R,\varepsilon} \sum_i (\rho_i + \sigma_i) + Ce^T \varepsilon \\
 \text{s.t.} \quad & \rho_i - \sigma_i = y_i(R - d_i) + \varepsilon_i, \quad d_i = \|x_i - O\|_2, \quad i = 1, \dots, n, \\
 & (\rho_i, \sigma_i, 1) \in S_3, \quad i = 1, \dots, n, \\
 & R \geq 0, \quad \varepsilon \geq 0.
 \end{aligned}$$

This problem is almost a Second Order Cone Program except that the equality constraints

$$\{d_i = \|x_i - O\|_2, i = 1, \dots, n\}$$

are nonlinear (which also makes the problem nonconvex). We use the first order Taylor expansion iteratively to approximate the nonlinear equalities by linear ones, which is detailed in the following algorithm in Section 4.2.



4.2. *An iterative algorithm to numerically solve radial DWD.* We consider applying the first order Taylor expansion iteratively to bypass the nonlinearity of the equality constraints and numerically solve Radial DWD:

*Initialization (Step 0):* Choose an initial center of the separating hypersphere and denote it as  $O^0$  (e.g., the mean or the coordinate-wise median of the +1 class training data), let the initial objective value be  $\text{Obj}^0 = -1$  (an arbitrary negative number).

*The iteration at Step  $k$ :*  $k \geq 1$ . Apply the first order Taylor expansion on  $d_i$  around  $O^{k-1}$ , that is,

$$\begin{aligned} d_i &= \|x_i - O\|_2 \\ &\approx \|x_i - O^{k-1}\|_2 + (\nabla_{O=O^{k-1}} \|x_i - O\|_2)^T (O - O^{k-1}) \\ &= \|x_i - O^{k-1}\|_2 - \frac{(x_i - O^{k-1})^T}{\|x_i - O^{k-1}\|_2} (O - O^{k-1}) := d'_i. \end{aligned}$$

Notice that  $d'_i$  is a linear function of  $O$ . By substituting  $d_i$  with  $d'_i$ , the optimization becomes a valid Second Order Cone Program and could be solved for  $O^k$  and  $R^k$  using SDPT3. Let  $\text{Obj}^k$  be the current objective value at step  $k$ .

*Stop:* if  $|\text{Obj}^k - \text{Obj}^{k-1}| < \epsilon$ , where  $\epsilon$  is a predetermined precision parameter.

At each step  $k$ , to approximate well the nonlinear terms by using the first order Taylor expansion, we further confine  $O^k$  in a neighborhood of  $O^{k-1}$  (the solution computed from the previous step) by adding one more constraint:  $\|O^k - O^{k-1}\|_2 \leq \delta_k$ , where  $\delta_k \in R_+$  is called the *step length* parameter. A small  $\delta_k$  guarantees the precision of the Taylor expansion but may slow down the computation. This additional constraint is a Second Order Cone constraint  $(\delta_k, O^k - O^{k-1}) \in S^{d+1}$  so that we still end up with a valid Second Order Cone Program at step  $k$ . In our current data analysis, we choose  $\epsilon = 10^{-4}$  and  $\delta_k = 10^{-3}$ . The choice of penalty  $C$  will be revisited after the discussion of Radial DWD optimality conditions in Section 4.3.

4.3. *The dual problem of radial DWD.* To gain more insights about Radial DWD optimization, it is useful to give the dual formulation of the Second Order Cone Program (at the  $k$ th step). Let  $w_i^{k-1} = \frac{x_i - O^{k-1}}{\|x_i - O^{k-1}\|_2} \in R^d$ ,  $d_i^{k-1} = \|x_i - O^{k-1}\|_2$  and they are functions of  $x_i$  (since  $O^{k-1}$  is computed from the previous step). After some algebra, we could formulate the dual program at step  $k$  as follows:

$$\text{Max}_z \sum_i y_i z_i d_i^{k-1} + \delta_k \left( - \left\| \sum_i y_i z_i w_i^{k-1} \right\|_2 \right) + 2 \sum_i \sqrt{z_i}$$

$$(4.3) \quad \text{s.t.} \quad 0 \leq z_i \leq C, \quad i = 1, \dots, n,$$

$$\sum_i y_i z_i \leq 0.$$

The primal and dual problems can be expressed more compactly in matrix-vector form. Keep all the defined notation unchanged and denote  $y$  as an  $n$ -vector of  $y_i$ ,  $z$  an  $n$ -vector of  $z_i$ ,  $\rho$  and  $\sigma$  the  $n$ -vectors of  $\rho_i$  and  $\sigma_i$ , respectively,  $Y$  an  $n$ -by- $n$  matrix with  $y_i$  on the diagonal. Additionally, let  $W_{k-1} = (w_1^{k-1}, \dots, w_n^{k-1}) \in R^{d \times n}$  with  $w_i^{k-1}$  defined above,  $\Delta_O^{k-1} = O - O^{k-1} \in R^d$  and  $d_{k-1} = (d_1^{k-1}, \dots, d_n^{k-1})^T \in R^n$ . Then, the primal-dual pair becomes

$$(4.4) \quad \begin{aligned} \text{(Primal)} \quad & \text{Min}_{\rho, \sigma, \Delta_O^{k-1}, R, \varepsilon} e^T \rho + e^T \sigma + C e^T \varepsilon \\ \text{s.t.} \quad & \sigma - \rho + Ry + YW_{k-1}^T \Delta_O^{k-1} + \varepsilon = Yd_{k-1}, \\ & (\delta_k, \Delta_O^{k-1}) \in S_{d+1}, \quad (\rho_i, \sigma_i, 1) \in S_3, \quad i = 1, \dots, n, \\ & R \geq 0, \varepsilon \geq 0, \end{aligned}$$

$$(4.5) \quad \begin{aligned} \text{(Dual)} \quad & \text{Max}_z d_{k-1}^T Yz + \delta_k (-\|W_{k-1} Yz\|_2) + 2e^T \sqrt{z} \\ \text{s.t.} \quad & 0 \leq z \leq Ce, \\ & y^T z \leq 0, \end{aligned}$$

where  $\sqrt{z}$  is a  $n$ -vector with  $\sqrt{z_i}$  as entries.

One can show the existence of strict feasible solutions to both the primal and dual problems. Since the primal and the dual are convex, it follows that the solution of the following optimality conditions are guaranteed to be optimal or, equivalently, the following equations are sufficient and necessary optimality conditions:

$$\begin{aligned} & \sigma - \rho + Ry + YW_{k-1}^T \Delta_O^{k-1} + \varepsilon = Yd_{k-1}, \\ & 0 < z \leq Ce, \quad \varepsilon \geq 0, \quad (Ce - z)^T \varepsilon = 0, \\ & R \geq 0, \quad y^T z \leq 0, \quad R(y^T z) = 0. \end{aligned}$$

Either  $W_{k-1} Yz = 0$  and  $\|O - O^{k-1}\| \leq \delta_k$ ,  
or  $\|O - O^{k-1}\|_2 = \delta_k (W_{k-1} Yz) / \|W_{k-1} Yz\|_2$ ,

$$\rho_i = \frac{z_i + 1}{2\sqrt{z_i}} \sigma_i = \frac{z_i - 1}{2\sqrt{z_i}} \quad \text{for all } i = 1, \dots, n.$$

It is important to note that the optimal radius is strictly positive in case the 2 training classes are separable and the penalty term  $C$  is large enough, which is shown in Theorem 1 [see the supplementary materials in Xiong, Dittmer and Marron (2015)]. If this is true, we could replace the condition

$\{R \geq 0, y^T z \leq 0, R(y^T z) = 0\}$  by  $\{R > 0, y^T z = 0\}$ . As one will see in Section 4.4, this condition gives an insight to the Radial DWD optimization. Besides, Theorem 1 also implies that the choice of the penalty parameter  $C$  should satisfy the following:  $C(d_i^{k-1})^2 > 1$  for all  $d_i, i \in \{i : y_i = -1\}$ .

Solving the primal/dual problem in an ultra high dimension may be inefficient. To deal with this issue, we first factor the data matrix  $X$  using a *QR decomposition*, for example,  $X = QU$  where  $Q \in R^{d \times n}$  has orthonormal columns and  $U \in R^{n \times n}$  is an upper triangular matrix. Then we solve the optimization problem by replacing  $X$  by  $U$  and call it a *reduced problem*. The reduced problem could be solved more efficiently because it shrinks the intrinsic dimension of the problem from  $d$  to the sample size  $n$ . Note that it is fairly easy to recover  $X$  from  $U$  once we solve the reduced problem. The reduced problem does not change the optimal solution or the optimal value, which is shown in Theorem 2 [see the supplementary materials in Xiong, Dittmer and Marron (2015)].

**4.4. Interpretation of the radial DWD dual problem.** Assume that the two classes are separable with a “proper” hypersphere (a hypersphere with nonzero radius  $R > 0$ ) so that  $y^T z = 0$  is obtained at optima. Notice that  $y^T z = 0$  implies  $e_+^T z_+ = e_-^T z_-$ , where  $z_+(z_-)$  is the subvector of  $z$  corresponding to the +1 class (-1 class) and  $e_+(e_-)$  the corresponding vector of ones. It makes sense to scale  $z$  such that  $e_+^T z_+ = e_-^T z_- = 1$ . We can write  $z$  as  $\eta z^*$ , where  $\eta$  is a positive scalar and  $z^*$  satisfies the additional scaling condition. By maximizing the dual objective function with respect to  $\eta$  for a fixed  $z$ , we find if  $-d_{k-1}^T Y z^* + \delta_k(\|W_{k-1} Y z^*\|_2) > 0$ ,

$$\sqrt{\hat{\eta}} = \frac{e^T \sqrt{z^*}}{-d_{k-1}^T Y z^* + \delta_k(\|W_{k-1} Y z^*\|_2)}.$$

Equivalently, the dual objective function becomes

$$(4.6) \quad \text{Max}_{z^*} \quad \frac{(e^T \sqrt{z^*})^2}{-d_{k-1}^T Y z^* + \delta_k(\|W_{k-1} Y z^*\|_2)}.$$

Moreover,

$$\begin{aligned} & d_{k-1}^T Y z^* - \delta_k(\|W_{k-1} Y z^*\|_2) \\ &= \left( \sum_{i \in P} d_i^{k-1} z_i^* - \sum_{i \in N} d_i^{k-1} z_i^* \right) - \delta_k \left\| \sum_{i \in P} \frac{x_i - O^{k-1}}{d_i^{k-1}} z_i^* - \sum_{i \in N} \frac{x_i - O^{k-1}}{d_i^{k-1}} z_i^* \right\|_2, \end{aligned}$$

where  $P$  is the index set of the +1 class, and  $N$  the index set of the -1 class.

Since  $\sum_{i \in P} z_i^* = 1$  with  $z_i^* \geq 0$ ,  $\sum_{i \in P} d_i^{k-1} z_i^*$  is a convex combination of  $d_i^{k-1}, i \in P$ , and it can be interpreted as an *average distance* from the current center of the separating sphere to the +1 class data points. A similar

interpretation applies for  $\sum_{i \in N} d_i^{k-1} z_i^*$ . When the two classes are separable (and  $R > 0$ ), the positive (negative) data points will be located inside (outside) the separating sphere so that the average distances of negative points are larger than that of the positive ones, which implies

$$\sum_{i \in P} d_i^{k-1} z_i^* - \sum_{i \in N} d_i^{k-1} z_i^* < 0 \quad \text{or} \quad -d_{k-1}^T Y z^* > 0.$$

From the above observation,  $-d_{k-1}^T Y z^* + \delta_k(\|W_{k-1} Y z^*\|_2) > 0$  is true when two classes are separable. Note that  $-d_{k-1}^T Y z^*$  is a measure of *separability* of the 2 classes and the bigger the absolute value, the bigger the separability. Meanwhile,  $w_i^{k-1} = \frac{x_i - O^{k-1}}{d_i^{k-1}}$  is a vector of unit Euclidean norm, pointing from the current center to each data point. Define the *centroid* of the +1 (−1) class as the convex combination of  $w_i^{k-1}$ ,  $i \in P$  (or  $i \in N$ , resp.) under weights  $z_i^*$ . Therefore,  $\sum_{i \in N} \frac{x_i - O^{k-1}}{d_i^{k-1}} z_i^* - \sum_{i \in P} \frac{x_i - O^{k-1}}{d_i^{k-1}} z_i^*$  is the vector pointing from the *centroid* of the +1 class to the *centroid* of the −1 class, and its Euclidean norm scaled by  $\delta_k$  is also a measure of separability. As a consequence, the whole denominator of (4.6) is positive and is a measure of separability of the 2 classes. To ensure optima, the dual problem minimizes the separability between classes divided by the square of the sum of the square roots of the convex weights.

Note that in some situations, the proportions of the 2 classes in the data set may not reflect the real proportions in a target population due to sampling bias, or the 2 classes are extremely unbalanced. The separating boundary tends to be closer to the class with smaller training sample size. In the case of biased sampling or unbalanced data, a weighted version of Radial DWD is more appropriate. Qiao et al. (2010) developed a weighting scheme to improve linear DWD and we follow the same line to set up weighted Radial DWD, by optimizing the following objective function:

$$(4.7) \quad \sum_i w(y_i) \left\{ \frac{1}{r_i} + C e^T \varepsilon \right\}$$

subject to the same set of constraints defined before. Note that  $w(y_i)$  is the *weight* associated with the  $i$ th training data point and it only depends on the class label  $y_i$ . In our data analysis we use  $w(+1) = \frac{n_-}{n_+ + n_-}$ ;  $w(-1) = \frac{n_+}{n_+ + n_-}$  as default. The above discussion about the Radial DWD optimization could be easily generalized to the case when weights are applied.

**5. Conclusion.** In this article we have proposed a nonlinear binary classifier, Radial DWD, for the virus hunting data analysis, where the virus positive class is surrounded by the negative class in very diverse radial directions. Because of the nonlinearity of classes, linear methods, including

MD, LASSO logistic regression, SVM and DWD, perform poorly with high classification error. Meanwhile, kernel SVM shows a very limited improvement over its linear counterpart in high dimensions. Since standard nonlinear methods, including kernel methods, require a type of “data richness,” that is not present in the virus hunting problem. In particular, they work well in situations (such as all the usual machine learning examples) where training data can be found in all of the various regions where the test data will appear. But in our particular data context, that completely breaks down, so all the classical nonlinear methods fare just as poorly as the linear ones. By using a much more appropriate spherical separating boundary, Radial DWD shows both low false positive and low false negative classification error. These are shown by real data analysis and simulation studies. Its computation through solving a sequence of Second Order Cone Programs is efficient, even with high-dimensional data.

We believe Radial DWD will be applicable in some settings beyond virus hunting. This will happen in classification contexts where there is one class with relatively small variation, and the other with much larger variation tending toward a number of quite divergent directions. For example, cancer is a disease of sometimes massive disruption of the genome, and these disruptions can go in many diverse directions, while the normal genome is far more stable. Another potential for this methodology comes in imaging bones and cartilage, where the normal population is relatively homogeneous, but severe wear and other types of abnormalities can go in many directions in the image space.

**Acknowledgments.** The authors would like to thank the Editor, Associate Editor and referees for their insightful, constructive comments.

## SUPPLEMENTARY MATERIAL

**Supplement to: “Virus hunting” using Radial Distance Weighted Discrimination.** (DOI: [10.1214/15-AOAS869SUPP](https://doi.org/10.1214/15-AOAS869SUPP); .pdf). In the supplementary materials, we first introduce some useful biology background for virus detection in Section 1, DNA alignment process in Section 2, and then discuss the insights of the Dirichlet distribution in Section 3. Real data examples and simulation studies are included in Sections 4 and 5, respectively. Theorems and proofs are in Section 6.

## REFERENCES

- ALIZADEH, F. and GOLDFARB, D. (2003). Second-order cone programming. *Math. Program.* **95** 3–51. [MR1971381](#)
- BURGES, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2** 955–974.

- FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. [MR2640659](#)
- GOLDSTEIN, D. B., ALLEN, A., KEEBLER, J., MARGULIES, E. H., PETROU, S., PETROVSKI, S. and SUNYAEV, S. (2013). Sequencing studies in human genetics: Design and interpretation. *Nat. Rev. Genet.* **14** 460–470.
- GRADA, A. and WEINBRECHT, K. (2013). Next-generation sequencing: Methodology and application. *J. Invest. Dermatol.* **133** e11.
- HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 427–444. [MR2155347](#)
- HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, New York. [MR2722294](#)
- JIANG, J., MARRON, J. S. and JIANG, X. (2009). Robust centroid based classification with minimum error rates for high dimension, low sample size data. *J. Statist. Plann. Inference* **139** 2571–2580. [MR2523649](#)
- JUNG, S. and MARRON, J. S. (2009). PCA consistency in high dimension, low sample size context. *Ann. Statist.* **37** 4104–4130. [MR2572454](#)
- LIU, Y., HAYES, D. N., NOBEL, A. and MARRON, J. S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *J. Amer. Statist. Assoc.* **103** 1281–1293. [MR2528840](#)
- MARRON, J. S., TODD, M. J. and AHN, J. (2007). Distance-weighted discrimination. *J. Amer. Statist. Assoc.* **102** 1267–1271. [MR2412548](#)
- METZKER, M. L. (2010). Sequencing technologies—the next generation. *Nat. Rev. Genet.* **11** 31–46.
- MWENIFUMBO, J. C. and MARRA, M. A. (2013). Cancer genome-sequencing study design. *Nat. Rev. Genet.* **14** 321–332.
- QIAO, X., ZHANG, H. H., LIU, Y., TODD, M. J. and MARRON, J. S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *J. Amer. Statist. Assoc.* **105** 401–414. [MR2656058](#)
- REHM, H. L. (2013). Disease-targeted sequencing: A cornerstone in the clinic. *Nat. Rev. Genet.* **14** 295–300.
- SCHÖLKOPF, B. and SMOLA, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.
- SHAWE-TAYLOR, J. and CRISTIANINI, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, Cambridge, MA.
- SHEN, D., SHEN, H. and MARRON, J. S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *J. Multivariate Anal.* **115** 317–333. [MR3004561](#)
- TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. [MR1379242](#)
- TUTUNCU, R. H., TOH, K. C. and TODD, M. J. (2001). SDPT3—a MATLAB software package for semidefinite-quadratic-linear programming. Available at <http://www.math.cmu.edu/users/reha/home.html>.
- VAPNIK, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. [MR1367965](#)
- World Health Organization WHO (2014). Middle East respiratory syndrome coronavirus (MERS-CoV) summary and literature update-as of 9 May 2014. Available at [http://www.who.int/csr/disease/coronavirus\\_infections/MERS\\_CoV\\_Update\\_09\\_May\\_2014.pdf](http://www.who.int/csr/disease/coronavirus_infections/MERS_CoV_Update_09_May_2014.pdf).
- XIONG, J., DITTMER, D. P. and MARRON, J. S. (2015). Supplement to: “Virus hunting” using Radial Distance Weighted Discrimination. DOI:[10.1214/15-AOAS869SUPP](https://doi.org/10.1214/15-AOAS869SUPP).

YATA, K. and AOSHIMA, M. (2013). PCA consistency for the power spiked model in high-dimensional settings. *J. Multivariate Anal.* **122** 334–354. [MR3189327](#)

J. XIONG

DEPARTMENT OF STATISTICS  
AND OPERATIONS RESEARCH  
UNIVERSITY OF NORTH CAROLINA  
AT CHAPEL HILL  
NORTH CAROLINA 27599-3260  
USA  
E-MAIL: [xiongj@unc.edu](mailto:xiongj@unc.edu)

D. P. DITTMER

LINEBERGER COMPREHENSIVE CANCER CENTER  
UNIVERSITY OF NORTH CAROLINA  
AT CHAPEL HILL  
450 WEST DRIVE, CB# 7295  
CHAPEL HILL, NORTH CAROLINA 27599-7295  
USA  
E-MAIL: [dirk\\_dittmer@med.unc.edu](mailto:dirk_dittmer@med.unc.edu)

J. S. MARRON

DEPARTMENT OF STATISTICS  
AND OPERATIONS RESEARCH  
UNIVERSITY OF NORTH CAROLINA  
AT CHAPEL HILL  
NORTH CAROLINA 27599-3260  
USA  
E-MAIL: [marron@unc.edu](mailto:marron@unc.edu)